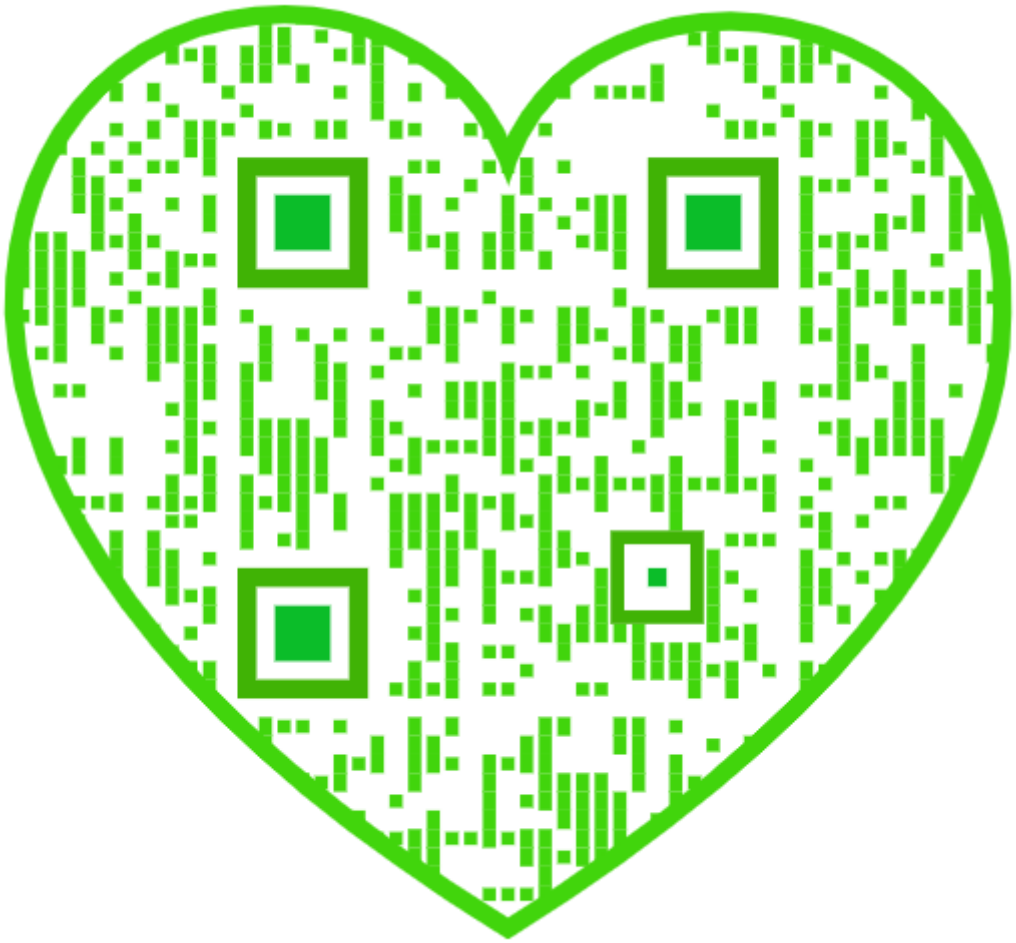
Master in Artificial Intelligence

Data Collection & Preprocessing II



Purpose

The purpose of the section is to help you learn how to collect and preprocess data to become a Successful Artificial Intelligence (AI) Engineer

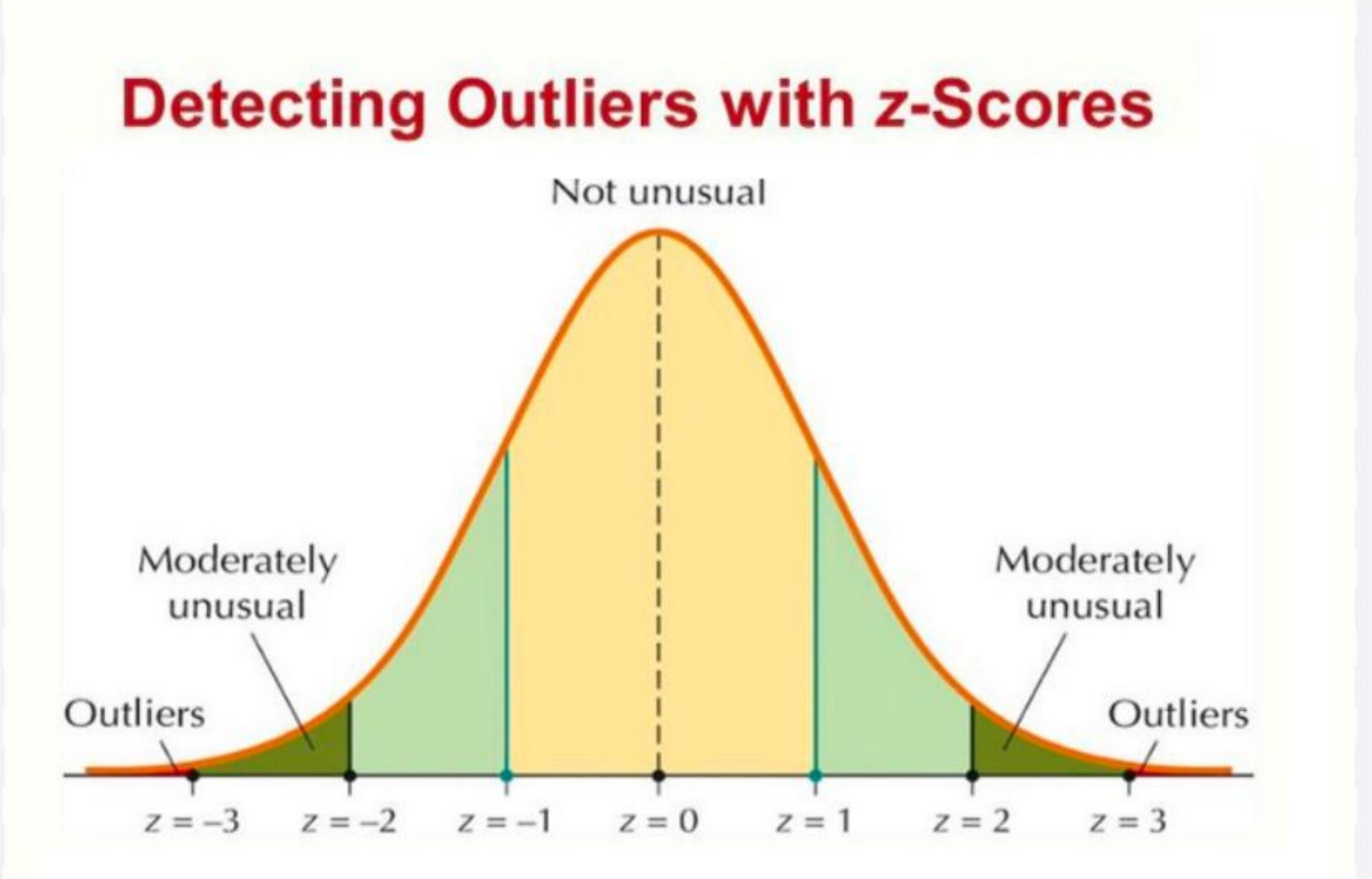
At the end of this lecture, you will learn the following

- **How to gather relevant data from various sources, ensure its quality, and preprocess it to make it suitable for analysis and modeling**

Standardizing data formats

	Values in Different Formats	Standardized Data
Dates	Dates in international document: • 12/31/2023 • 21-12-2023 • 2023/11/15	ISO 8601 Format: YYYY/MM/DD • 2023/12/31 • 2023/12/21 • 2023/11/15
Measurement Units	Different weight formats: • 150 pounds • 64 oz • 11 stone	Standard Format: KG • 68.039 kg • 1.814 kg • 69.853 kg
Language Translation	Phrases from news articles in different languages	All content translated to English for consistency

Dealing with outliers- What are Outliers?



Dealing with outliers

Robust Methods for Handling Outliers

Trimming

Winsorization

Robust regression

Robust standard errors

Feature Engineering

Creating new features

+

Transforming existing features

Enhance the predictive power of the data for modeling

Encoding categorical variables

CATEGORICAL VARIABLES

DEFINITION

Categorical variables represent data that can be divided into multiple categories but cannot be ordered or measured. Each category can be identified by a distinct label, and data points are allocated to these categories based on qualitative properties. These variables can further be broken down into ordinal, binary, and nominal variables.

EXAMPLES

- **Hair Color (Nominal):** categories include "blonde", "brunette", "black", and "red".
- **Has a Pet (Binary):** You either have a pet or you don't, making this a binary variable.
- **Ranking (Ordinal):** positions like "first", "second" & "third" represent an ordinal variable. The positions clearly depict a ranking order.

HELPFULPROFESSOR.COM

Converting categorical variables into numerical representations

Person	Marital status
xxx	Single
yyy	Married
zzz	Divorcee

Categorical to binary

Person	Single	Married	Divorcee
xxx	1	0	0
yyy	0	1	0
zzz	0	0	1

Binary to numeric

Person	Martital status
xxx	4
yyy	2
zzz	1

Converting categorical variables into numerical representations

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Converting categorical variables into numerical representations

FRUITS	YEAR 1	YEAR 2
APPLE	80	100
MANGO	50	100

↓

FRUITS	YEAR 1	YEAR 2
90	80	100
75	50	100

TARGET ENCODING

Steering Categories with Outcomes

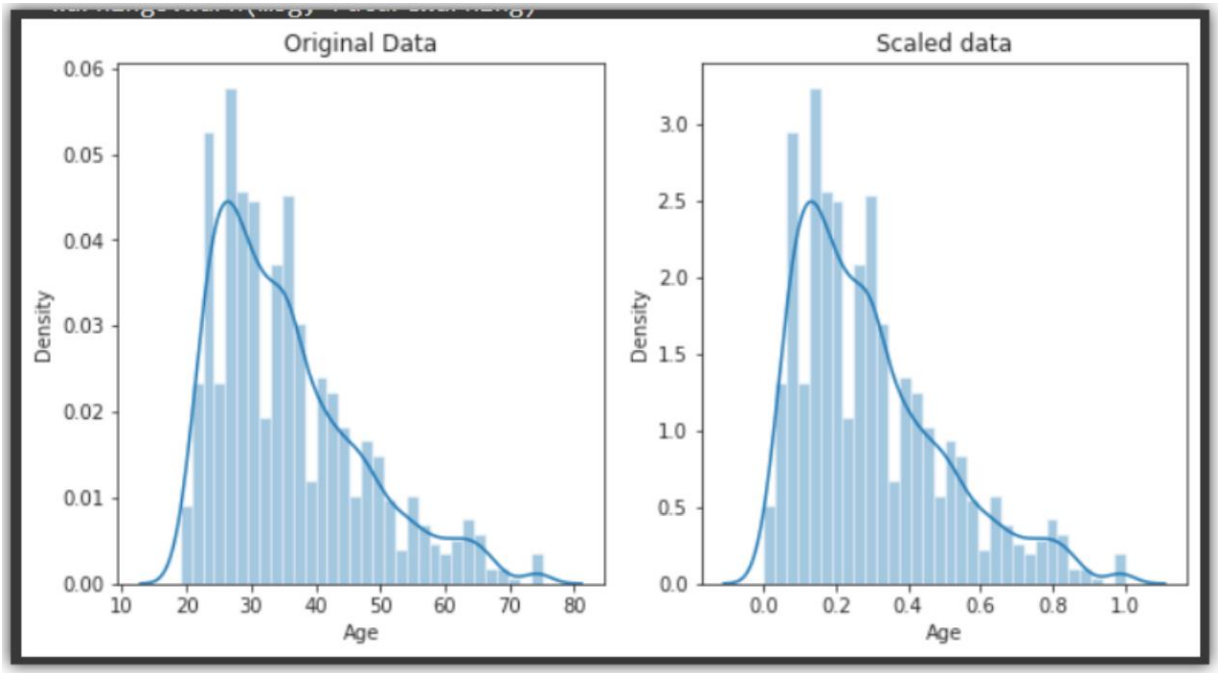
Scaling and normalization

1. Simple Feature Scaling:

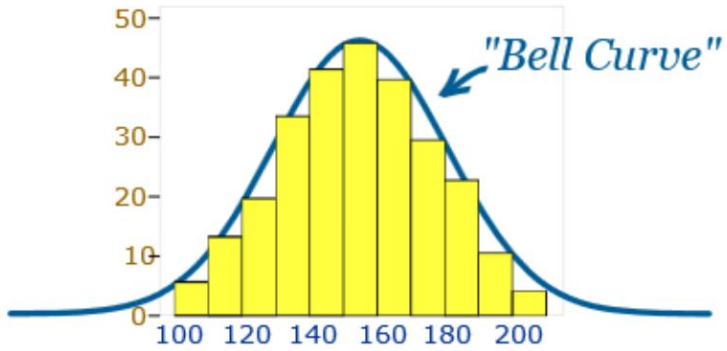
$$Xnew = \frac{Xold}{Xmax}$$

2. Min-Max Scaling:

$$Xnew = \frac{Xold - Xmin}{Xmax - Xmin}$$

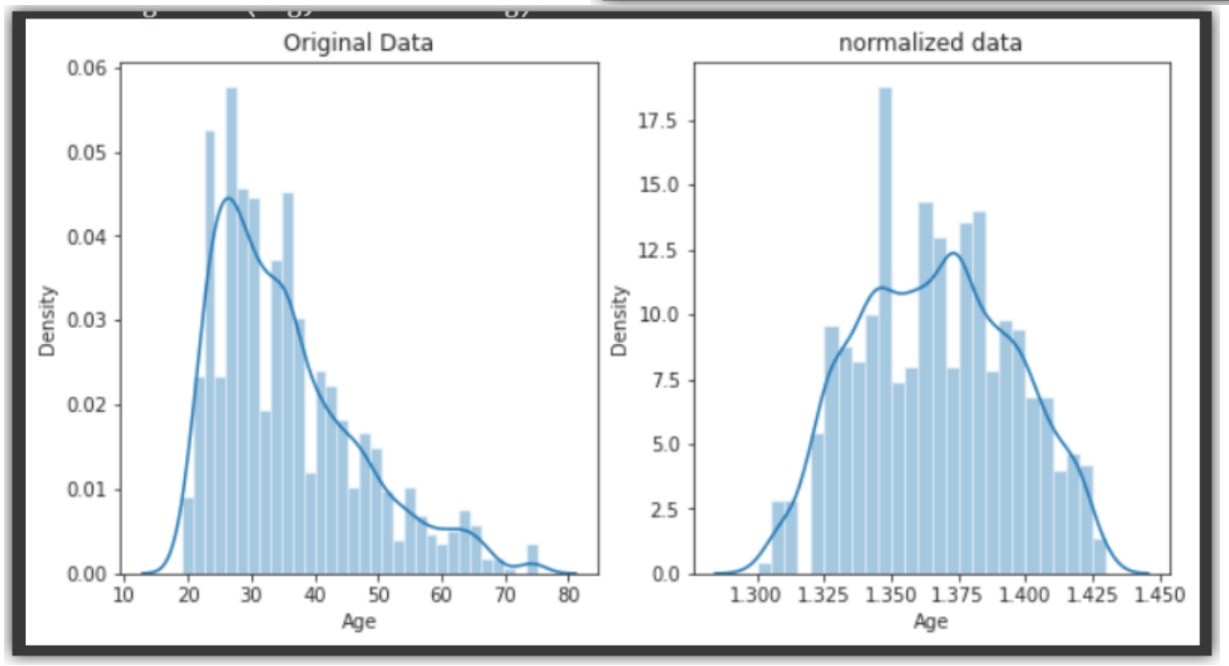


Scaling and normalization



$$Xnew = \frac{Xold - mean}{STD(sigma)}$$

$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ (y_t^{\lambda} - 1)/\lambda & \text{otherwise.} \end{cases}$$



What is next?

Feature extraction

Master in Artificial Intelligence

Data Collection & Preprocessing II